Démotivation des dérivés morphologiques

Cécile Fabre*, **Nabil Hathout***, **Lydia-Mai Ho-Dac***
en collaboration avec
Alizée Lombard (Université de Fribourg)
Marine Wauquier (LATTICE & Université de Paris 3)
Richard Huyghe (Université de Fribourg)

*CLLE, CNRS & Université Toulouse Jean Jaurès

Thématiques actuelles de la recherche en TAL
24 octobre 2022

Contexte et objectifs du projet

Partenariat Hubert Curien - Germaine de Staël

- Exploiter les mesures distributionnelles dans la recherche en sémantique lexicale :
  - délimitation de catégories nominales (Huyghe et Wauquier 2020)
  - analyse morphosémantique (Wauquier 2020, Wauquier et al. 2020)
- Interroger la portée de ces approches pour construire des représentations sémantiques pertinentes pour la linguistique
- Bénéficier de l'apport complémentaire des méthodes expérimentales en interrogeant conjointement les perceptions et les usages
- Au-delà des aspects méthodologiques, répondre à des questions linguistiques précises.

(...) most work is directed at showing that distributional semantics can model derivational morphology, rather than tackling more specific linguistic question." (Boleda 2020, p.16)

Démotivation morphosémantique

- La lexicalisation des lexèmes dérivés (= leur inclusion dans le lexique) peut s'accompagner d'une perte de compositionnalité (Hilpert 2020).

- La démotivation morphosémantique est la perte de la relation sémantique entre un dérivé et sa base. Le sens du lexème devient opaque.

- La démotivation des dérivés lexicalisés est variable, graduelle (Roché 2004) :

Dérivés motivés	*abattre* / *abattoir*
Dérivés partiellement motivés	*trotter* / *trottoir*
Dérivés démotivés	*serrer* / *serrure*

Mesurer la démotivation

- L'étude porte sur la transparence sémantique dans des couples formés d'un verbe et d'un nom du français morphologiquement apparentés.

- Deux objectifs :
  1. concevoir des méthodes permettant d'identifier le phénomène de démotivation ;
  2. en explorer la nature graduelle.

- Combiner deux méthodes habituellement utilisées pour estimer la transparence sémantique d'un lexème :
  - méthode expérimentale fondée sur les jugements de proximité exprimés par des locuteurs (Gagné *et al.* 2016, Creemers *et al.* 2020) ;
  - méthode computationnelle fondée sur des scores de similarité distributionnelle (Marelli & Baroni 2015, Varvara *et al.* 2021).

Mesurer la démotivation

- On fait l'hypothèse que les mesures expérimentales et distributionnelles de la démotivation vont converger
- On cherche ensuite à identifier les paramètres qui jouent sur la démotivation, et en particulier le rôle des suffixes

Sélection des données

- Définition de 3 conditions permettant d'observer la gradabilité de la démotivation :

  C1 = démotivation totale (*partir* / *partage*)
  C2 = démotivation partielle (*mouiller* / *mouillette*)
  C3 = motivation totale (*danser* / *danseur*)

- Les couples sélectionnés sont tels que:
  - Le nom peut être analysé formellement comme un dérivé construit par suffixation à partir du verbe.
  - Une relation morphosémantique entre le nom et le verbe est attestée historiquement.
  - On exclut par exemple : *crever* / *crevette*

Sélection des données : les couples C1

- Liste de 16 suffixes identifiés comme supports de création de noms déverbaux
- Compilation de larges listes de noms à partir de lexiques (TLFi, GLàFF, Anagrimes) et de concordanciers (FRCOW16A)
- Point de départ : les couples C1
  - Condition la plus rare → identifier un ensemble le plus large possible de couples démotivés.
  - Répartition des recherches par suffixe puis adjudication

26 paires C1

8 suffixes avec au moins 2 occurrences : -ade, -age, -ance, -et, -ette, -eur, -oir, -ure

Sélection des données : les couples C2 et C3

- Répétition de la procédure pour C2 et C3
- Avec des contraintes sur la sélection des exemplaires :
  - 26 couples par condition,
  - Le jeu de données contient le même nombre de couples C1, C2, C3,
  - Les suffixes sont présents dans les mêmes proportions pour les trois conditions,
  - Les couples sont comparables en termes de longueur, de fréquence et de types sémantiques.
- Exemples :
  - pour *-oir*, des noms dénotant des lieux et des instruments sont présents dans les 3 classes,
  - le couple C3 *passer* / *passage* a été préféré à *échantillonner* / *échantillonnage* sur la base du couple C1 *partir* / *partage*

Exemples

	C1		C2		C3	
suffixe	**verbe**	**nom**	**verbe**	**nom**	**verbe**	**nom**
-ade	bouter	boutade	tailler	taillade	brimer	brimade
-age	partir	partage	taper	tapage	passer	passage
-ance	croire	créance	ordonner	ordonnance	attirer	attirance
-et	cacher	cachet	fumer	fumet	jouer	jouet
-ette	éprouver	éprouvette	mouiller	mouillette	sonner	sonnette
-eur	procurer	procureur	synthétiser	synthétiseur	danser	danseur
-oir	couler	couloir	trotter	trottoir	abattre	abattoir
-ure	serrer	serrure	fournir	fourniture	déchirer	déchirure

Difficulté

Mesurer le rôle des suffixes avec un échantillon de données aussi réduit sera difficile

Dispositif expérimental

Principe

- Les couples verbe-nom sont présentés à des locuteurs natifs du français dans un questionnaire en ligne.
- Les locuteurs doivent estimer la proximité sémantique entre les deux mots.

Hypothèse

Les couples démotivés (C1) devraient être jugés comme ayant une proximité sémantique plus faible que les couples motivés (C3).
Les jugements attendus pour les couples partiellement motivés (C2) devraient être intermédiaires.

Dispositif expérimental

Tâche

D'après vous, à quel point les sens du nom ***conservatoire*** et du verbe ***conserver*** sont-ils proches ?

0 - sans rapport
6 - extrêmement proches

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6

○ Je ne connais pas un de ces mots

Dispositif expérimental

Participants

- 411 étudiants de licence de l'Université Toulouse Jean Jaurès ont participé à l'enquête.
- 39 stimuli par questionnaire

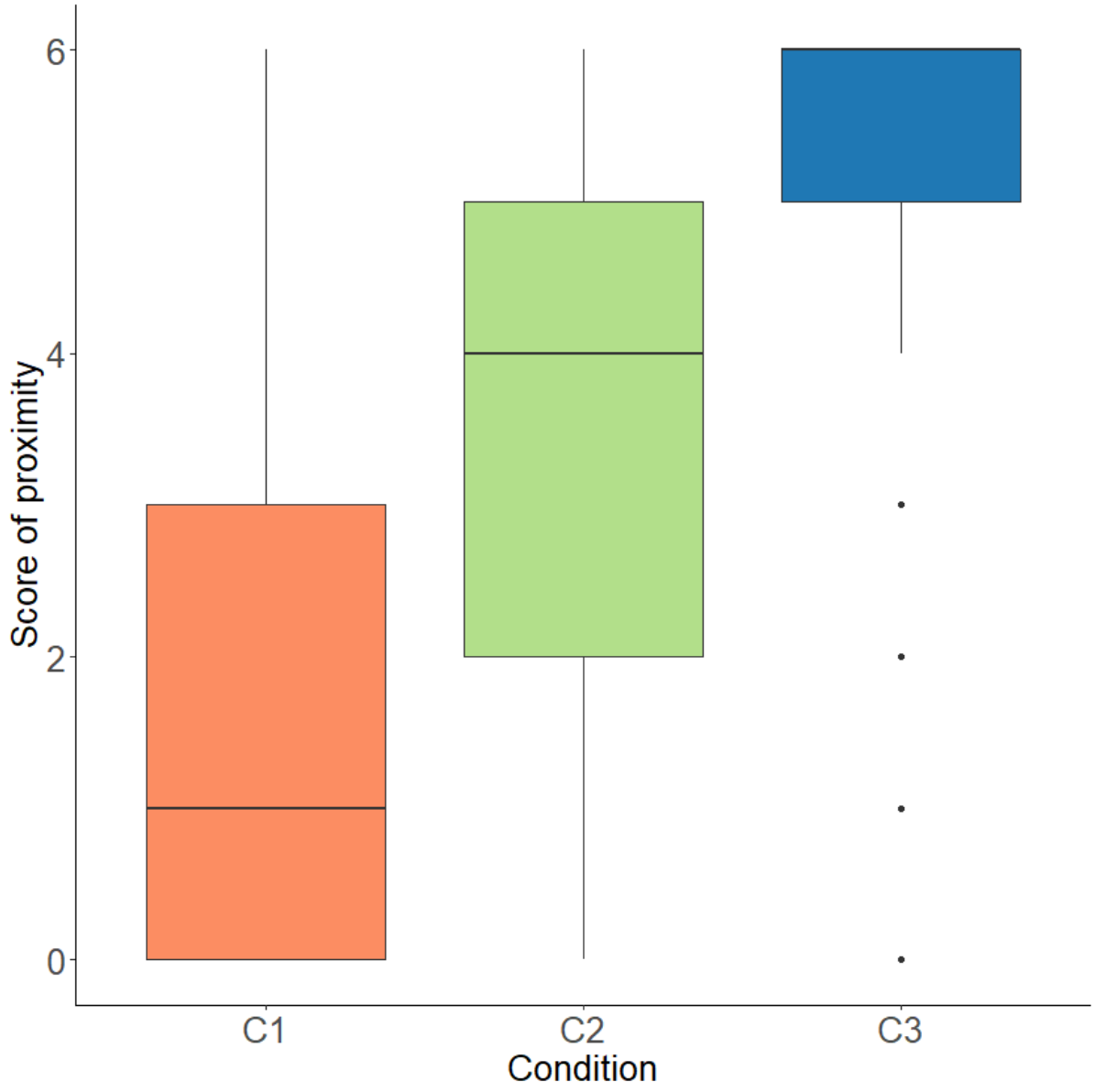
Filtrage des réponses

- Contraintes sur les participants :
  - Francophones natifs
  - Âge compris entre 17 et 25 ans ; moyenne = 19.4
- Contraintes sur les réponses :
  - Elimination des réponses présentant des temps de réponse atypiques

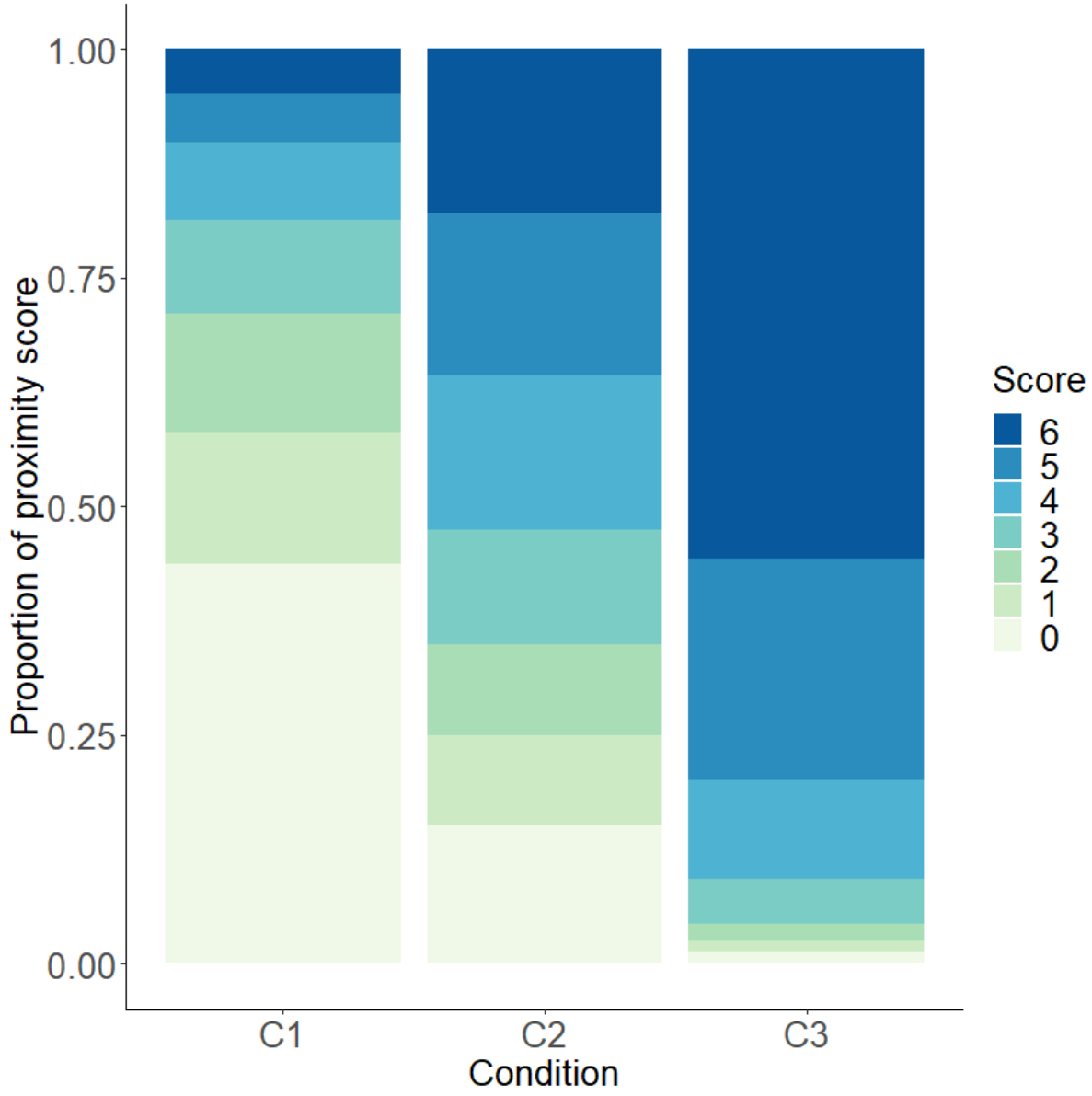
Nombre total d'observations par catégories

3210 pour C1, 3401 pour C2, 3378 pour C3

Résultats - scores de proximité par catégorie



Résultats - scores de proximité par catégorie



Résultats - vérification de l'hypothèse

- On cherche à vérifier statistiquement si les catégories C1-C2-C3 ont un impact sur le jugement des locuteurs
- Calcul d'une régression logistique ordinale
- Effet significatif de la condition C1-C2-C3 sur le score de proximité expérimental avéré $p < 2.2 \times 10^{-16}$.
- Effet non significatif de la fréquence du verbe et du nom.

Conclusion provisoire

Les résultats montrent que les jugements des participants et les intuitions des experts sont consistants.

➜ Les locuteurs sont sensibles à la gradation établie par les experts.

Résultats - étude de la variation interne aux catégories

- Étude des variations à l'intérieur des catégories
- Moyenne et déviation standard du score de proximité moyen des couples verbe-nom pour les 3 conditions

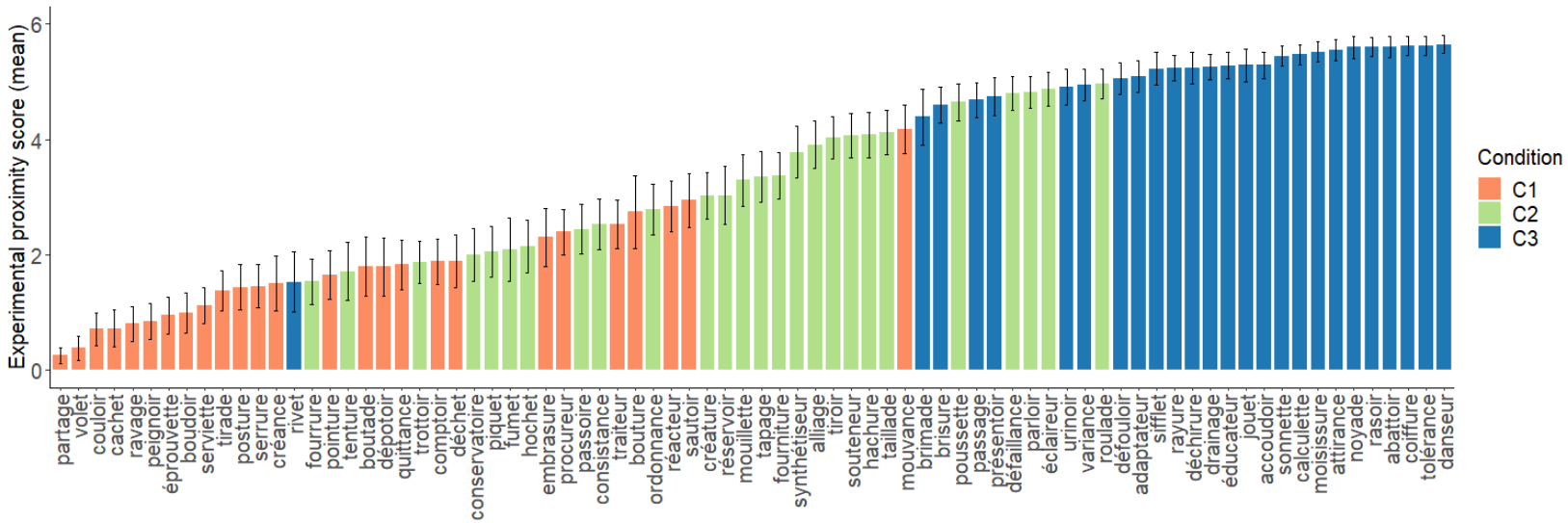
Cat.	N	Mean	SD
C1	26	1.67	1.61
C2	26	3.28	1.75
C3	26	5.09	1.06

Confirmation

la catégorie C3 est la plus homogène, C2 la moins homogène.

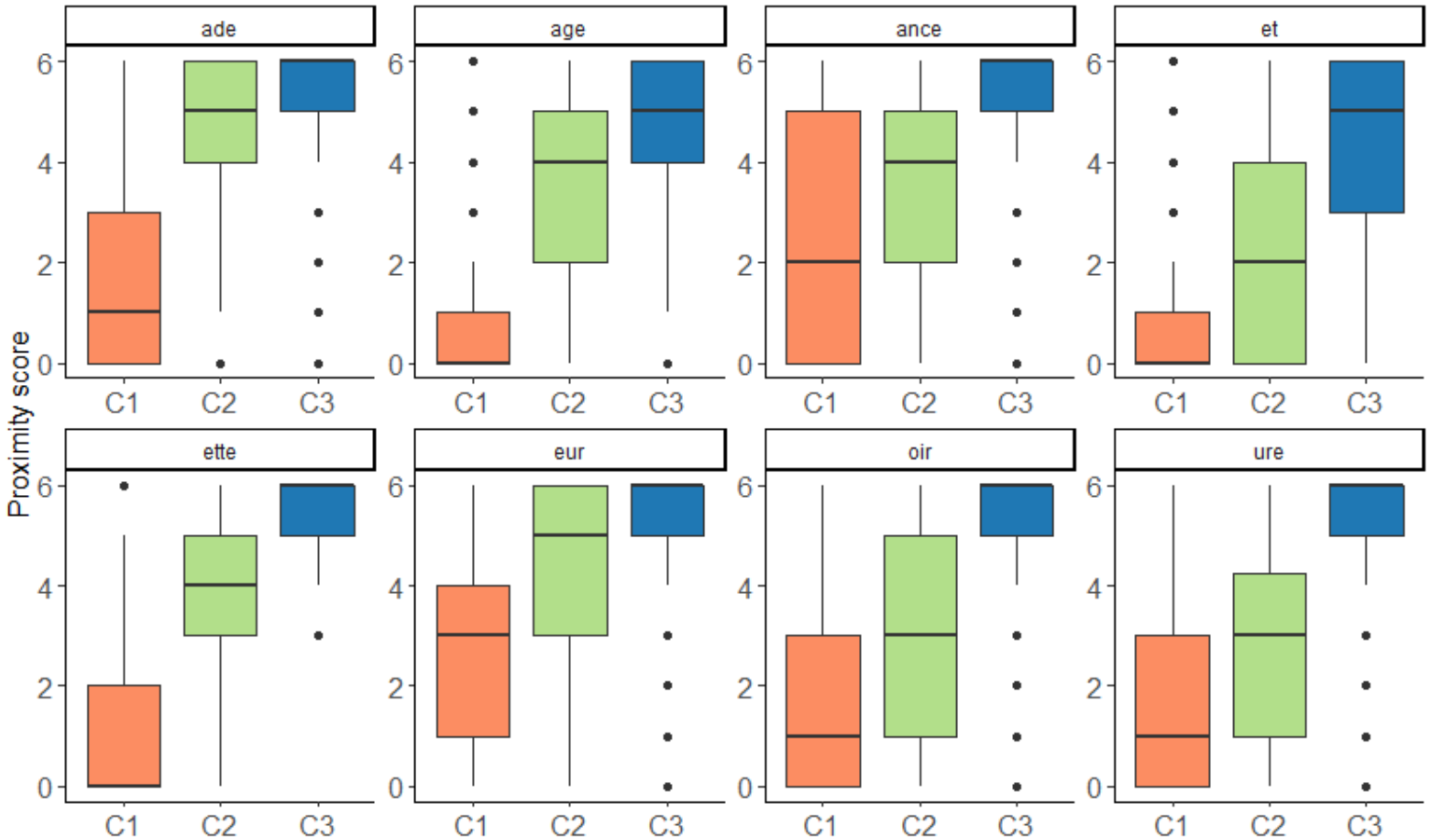
Résultats - étude de la variation par couple

Score moyen par couple verbe-nom :



Résultats - étude de la variation par suffixe

Distribution du score expérimental par suffixe



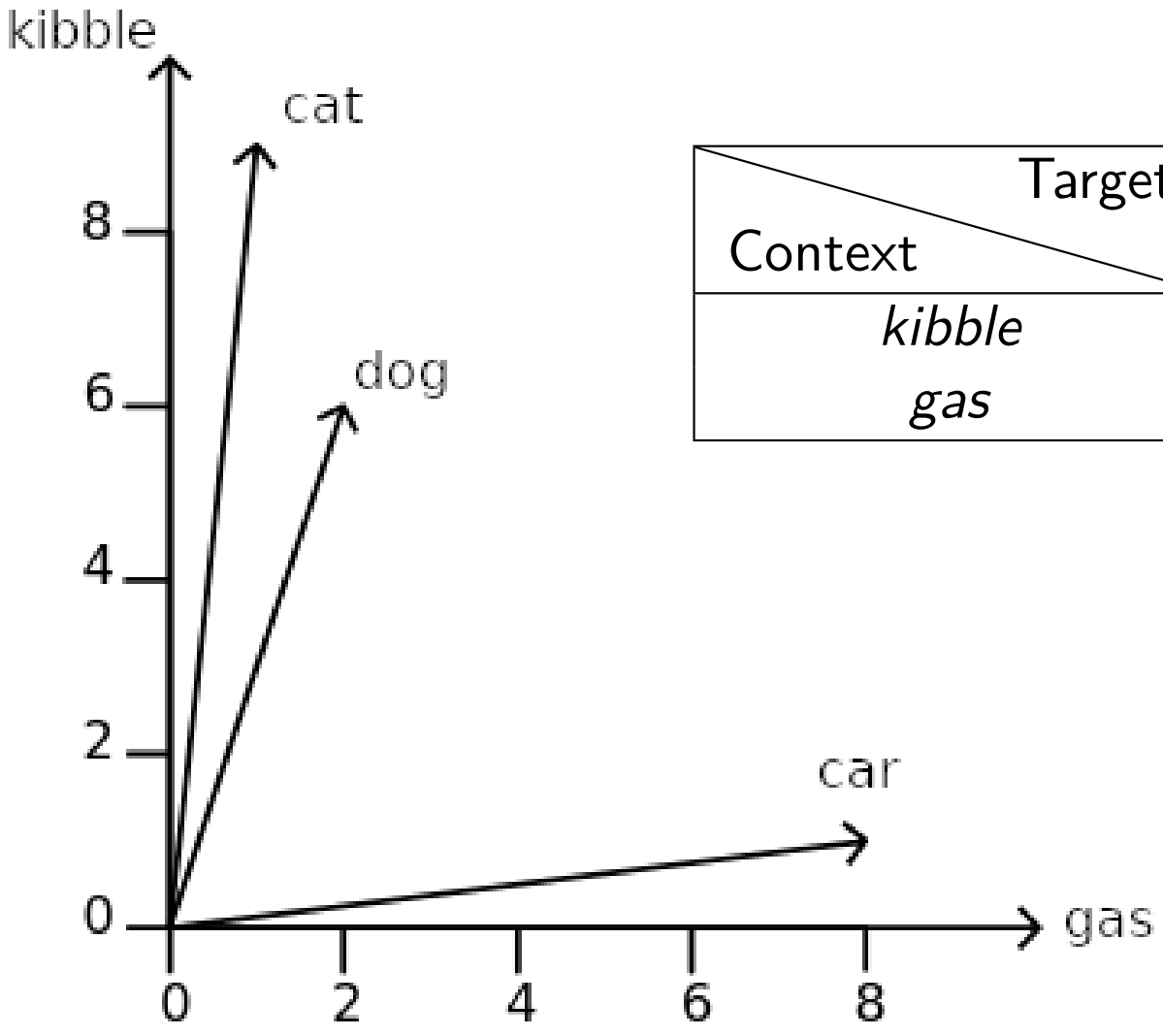
Rappels sur la sémantique distributionnelle

La sémantique distributionnelle s'appuie sur l'hypothèse que les mots qui ont des distributions similaires (i.e. qui apparaissent dans des contextes similaires) ont des sens similaires.

Le sens des mots dans un corpus est représenté par des vecteurs calculés à partir de leurs cooccurrences.

La similarité sémantique de deux mots est estimée par la proximité des vecteurs qui les représentent sur une échelle allant de 0 à 1.

Rappels sur la sémantique distributionnelle



Context \ Target	*dog*	*cat*	*car*
kibble	6	9	1
gas	2	1	8

Méthode

Nous avons calculé des *word embeddings* (plongements lexicaux) en utilisant word2vec (Mikolov *et al.* 2013) à partir du corpus français Wikipédia (900 millions de mots) lemmatisé en utilisant TALISMANE (Urieli 2013).

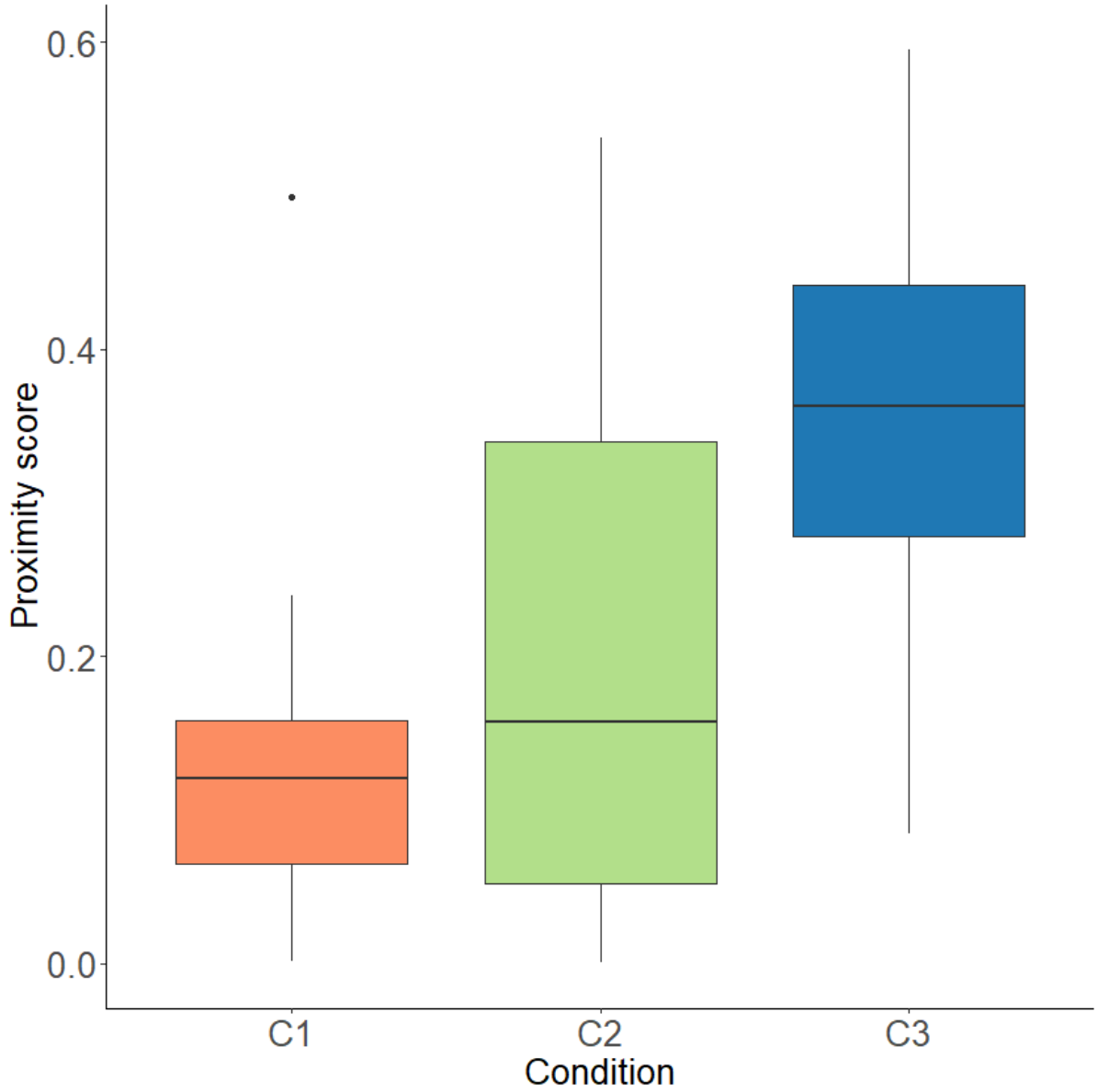
- Les représentations vectorielles que nous utilisons sont construites en concaténant 5 espaces vectoriels sémantiques construits de la même manière à partir du même corpus.
- Les hyperparamètres sont : CBOW, Negative Sampling, 100 dimensions, fréquence minimale $= 5$, fenêtre $= 5$

La similarité distributionnelle de chacun des couples verbe-nom du jeu de donnée est estimée par le cosinus des vecteurs des deux mots

Hypothèse

- Le score cosinus devrait être élevé pour les couples C3 (proche de 1) et faible pour les couples C1 (proche de 0)
- Les couples C2 devraient avoir des scores cosinus intermédiaires

Résultats



Résultats

Moyenne et déviation standard des scores de proximité distributionnels

Category	N	Mean	SD
C1	26	0.129	0.099
C2	26	0.206	0.181
C3	26	0.351	0.151

Les couples C3 ont en moyenne un score cosinus plus élevé que les couples C1 et C2.

La différence entre les trois conditions est significative ($p = 2.21 \times 10^5$).

La comparaison 2 à 2 des trois conditions montre que la différence est significative pour les couples C1 et C3, C2 et C3.
La différence n'est pas significative pour les couples C1 et C2.

La distribution rend compte des différences entre les trois niveaux de démotivation morphosémantique

Croisement des résultats

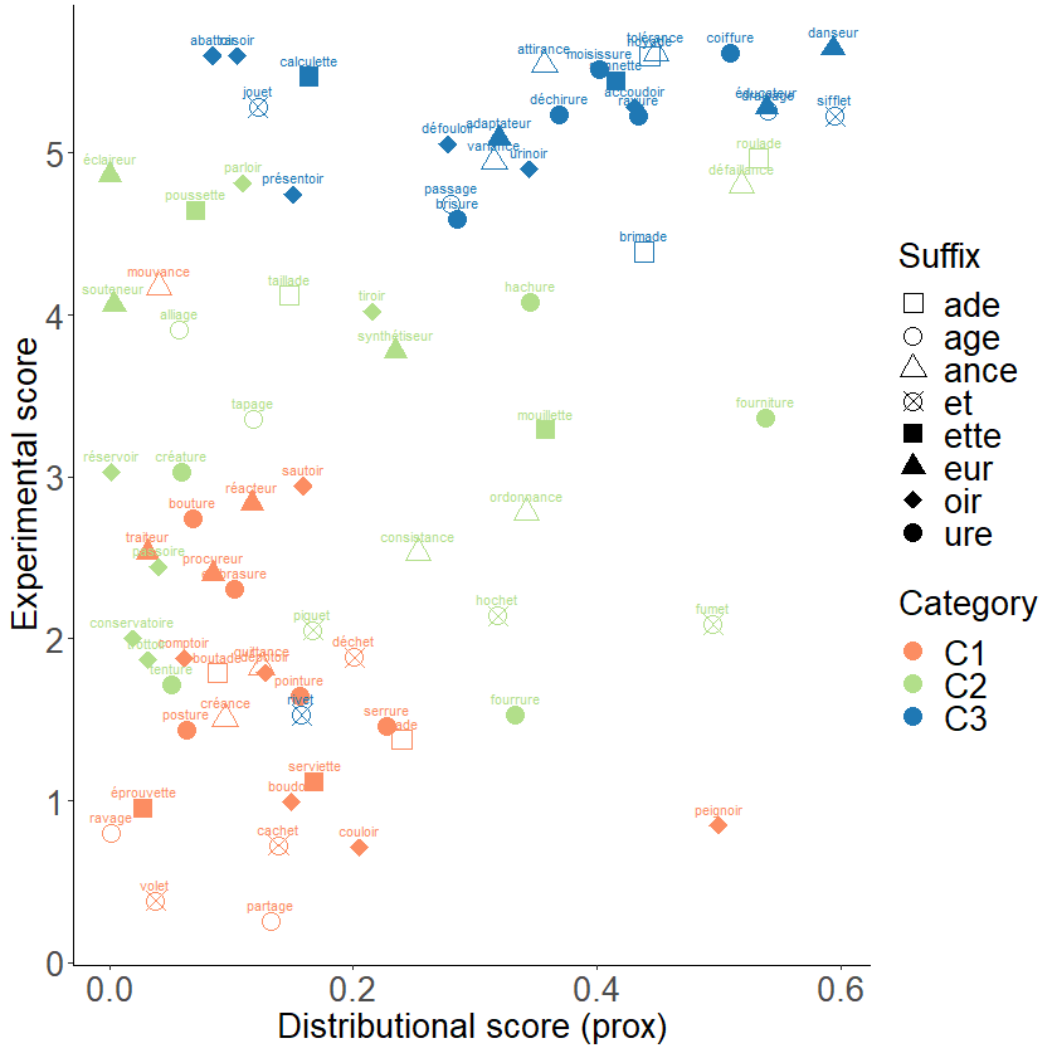
Le score distributionnel est-il un bon prédicteur du score expérimental ?

- Nous avons réalisé une régression ordinale mixte dans laquelle le score expérimental est une fonction du score distributionnel qui a permis d'obtenir une p-value de 1.9×10^{-6}

Effet	**Estimate**	**SE**	***z***	***p***
Prox	6.096	1.279	4.766	1.88×10^{-6}

Comparaison des deux approches

Score expérimental (1 à 6) et distributionnel (0 à 1) des couples verbe-nom.



Convergence

Les deux méthodes convergent globalement :

- C1 en bas à gauche ;
- C3 en haut à droite

Différences

Tendances

Les C3 (bleu) sont en haut : les humains perçoivent parfaitement la motivation.

Les C1 (orange) sont à gauche : les représentations distributionnelles rendent compte de la démotivation.

Différences

Les jugements humains (sur l'axe vertical) tendent à trouver des similarités pour les couples verbe-nom distants (beaucoup de vert en haut du graphique):

- *éclaireur* - *éclairer* sont sémantiquement distants, mais les locuteurs leur ont attribué un score moyen de 4.86
- *mouvance* - *mouvoir* sont sémantiquement très distants, mais les locuteurs leur ont attribué un score moyen de 4.17

Différences

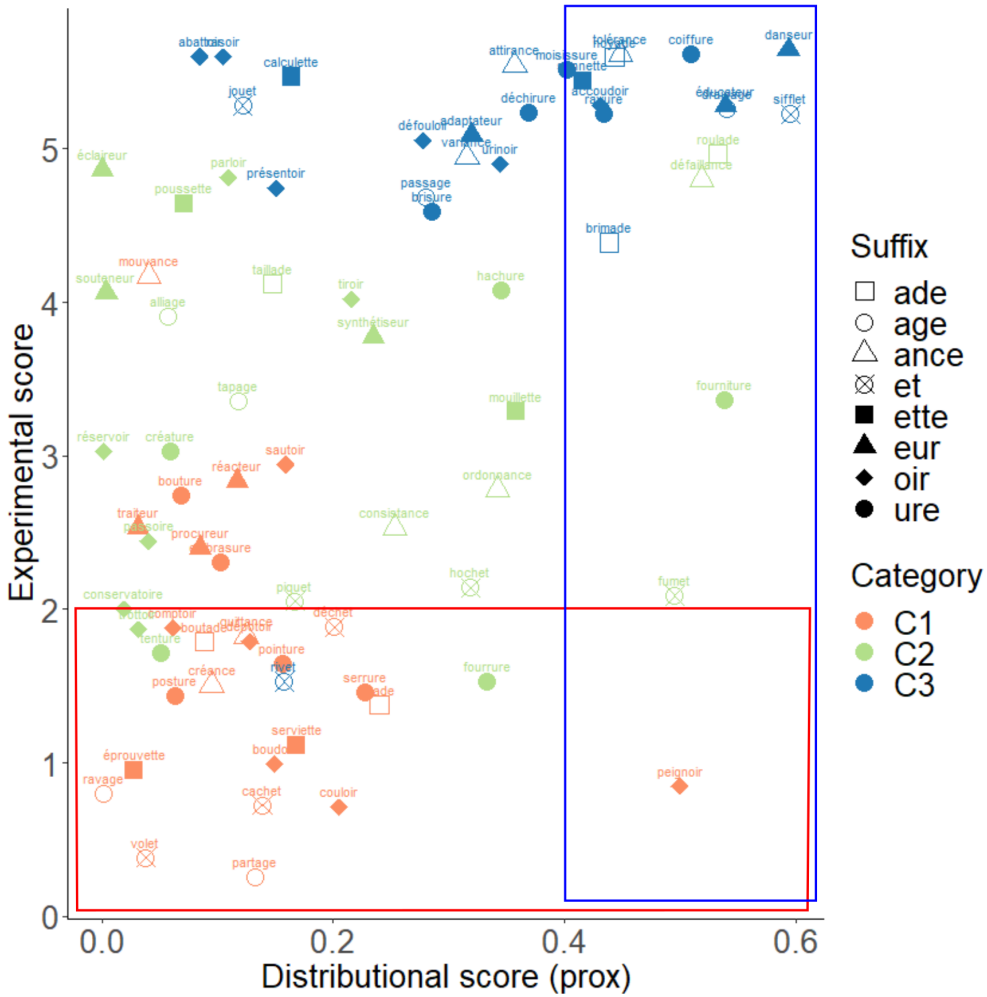
Les scores distributionnels (sur l'axe horizontal) tendent à considérer comme distants des couples motivés:

- la relation *abattoir* - *abattre* est transparente, mais les représentations distributionnelles des deux mots sont distantes (score = 0.08)
- la relation *jouet* - *jouer* est transparente, mais les représentations distributionnelles des deux mots sont distantes (score = 0.12)

Les scores distributionnels peuvent être affectés par des facteurs sans lien avec la motivation :

- *peignoir* - *peigner* ont un score de 0.84 parce qu'ils co-occurrent avec d'autres mots associés à la salle de bain.

Fiabilité



Les noms des couples jugés distants par les locuteurs ont de bonnes chances d'être démotivés

Les noms des couples dont les représentations distributionnelles sont proches ont de bonnes chances d'être motivés

Complémentarité

1. Confirmation du fait que les deux méthodes mettent en évidence deux clusters remarquables (C1 et C3)
2. La convergence des deux méthodes confirme leur fiabilité
   - Une forte similarité distributionnelle est un indice fort de motivation sémantique
   - Une proximité expérimentale faible est une indice fort de démotivation sémantique

Limites des deux méthodes

L'expérience fait apparaître certaines limites des deux méthodes.

1. Le coût de l'expérimentation est relativement élevé. Il empêche de la répéter fréquemment.

2. La méthode distributionnelle exploite un modèle générique tandis que la méthode expérimentale utilise un dispositif calibré pour répondre aux besoins de l'étude.

3. La taille du jeu de test est adaptée à l'expérience psycholinguistique mais elle est trop petite pour l'expérience distributionnelle.

4. La méthode distributionnelle ne dispose pas de représentation de la relation entre le verbe et le nom.

5. Ni la méthode expérimentale, ni la méthode distributionnelle ne testent la compositionnalité sémantique du sens du dérivé. L'instruction sémantique du suffixe n'est pas prise en compte.

Une caractéristique mesurable et gradable

Les méthodes expérimentales et distributionnelles peuvent être considérées comme complémentaires.

- Des preuves linguistiques (*linguistic evidence*) de type différent : intuitions des locuteurs vs usages.
- Score expérimental : une mesure plus fine de la démotivation, mais coûteuse et difficile à répliquer à grande échelle.
- Score distributionnel : une mesure automatisable, mais qui isole insuffisamment la propriété par rapport à d'autres (polysémie, partage de traits sémantiques).

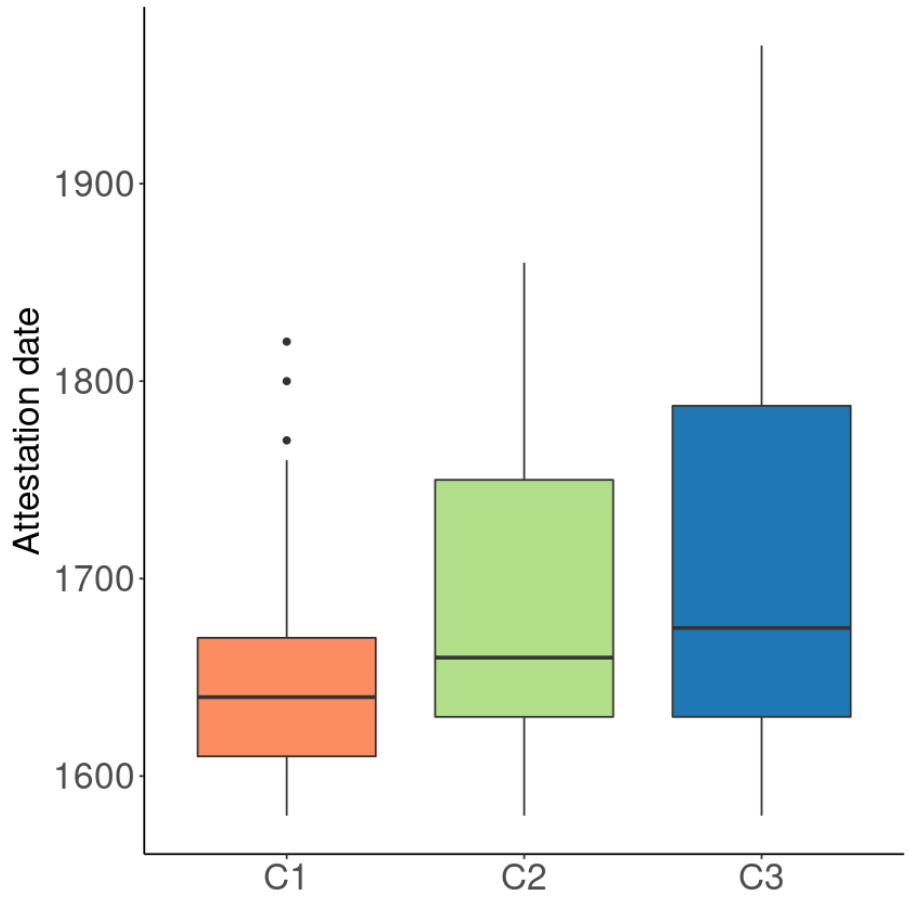
Les résultats mettent l'accent sur la gradabilité du phénomène d'un point de vue psycholinguistique et distributionnel.

- Les couples motivés et démotivés se positionnent aux deux extrémités de l'échelle.
- Les couples partiellement motivés sont répartis sur la totalité de l'échelle de similarité.

Perspectives - dimension diachronique de la démotivation

- Adaptation de l'hypothèse de (Baayen 1993) : plus un mot est fréquent, plus il est opaque.
- Hypothèse traduite en termes d'empan temporel : plus un mot est ancien, plus il est opaque.
- Utilisation de Google Ngrams (Bonami et Thuilier 2019).
- Les différences entre les dates d'attestation diffèrent significativement entre les 3 groupes (test de Kruskal-Wallis).
- Confirmation avec le score expérimental, mais pas avec le score distributionnel.

Perspectives - dimension diachronique de la démotivation



Perspectives

[✗] L'étude est évidemment limitée par la taille des échantillons.

- Etendre le matériel linguistique
  - Diversité morphologique : couples nom-nom (*poisson-poissonnier*), adjectif-nom (*haut-hauteur*), etc.
  - Contrôle plus strict sur les propriétés morphologiques (types sémantiques, ambiguité)
- Etudier l'effet des affixes et de la productivité sur la démotivation
  - Est-ce que le niveau de démotivation est plus important pour certains affixes relativement à d'autres ?
  - La démotivation est-elle corrélée au niveau de la productivité des affixes ?
- Explorer d'autres mesures de la transparence sémantique :
  - en psycholinguistique (effets d'amorce)
  - en sémantique distributionnelle (inclusion distributionnelle)

Perspectives

À paraître

Lombard, A., Wauquier, M., Fabre, C., Hathout, N., Ho-Dac, L.-.M & Huyghe, R. (2022), Evaluating morphosemantic demotivation through experimental and distributional methods, *Linguistic Investigationes*, 45:1.

Références

Baayen, 1993. On frequency, transparency and productivity. In *Yearbook of morphology 1992*, p. 181–208. Springer.

Bauer, Laurie. 1983. *English word-formation.* Cambridge University Press.

Blank, Andreas. 2001. *Pathways of lexicalization* 1596–1608. De Gruyter Mouton.

Boleda, G. (2020). Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6, 213-234.

Bonami O. Thuilier J. (2019). A statistical approach to rivalry in lexeme formation: French -iser and -ifier. *Word structure*, 12(1), 4–41.

Brinton, Laurel J. & Elizabeth Closs Traugott. 2005. *Lexicalization and language change*. Cambridge University Press.

Corbin, Danielle. 1987. *Morphologie dérivationnelle et structuration du lexique*. Mouton De Gruyter.

Creemers, Ava, Amy Goodwin Davies, Robert J. Wilder, Meredith Tamminga & DavidEmbick. 2020. Opacity, transparency, and

Références

morphological priming: A study of prefixed verbs in dutch. *Journal of Memory and Language* 110.

Gagné, Christina L., Thomas L. Spalding & Kelly A. Nisbet. 2017. Processing english compounds: Investigating semantic transparency. *SKASE Journal of Theoretical Linguistics* 13(2). 2–22.

Hilpert, Martin. 2019. *Lexicalization in morphology*. Oxford University Press.

Huyghe, R. & Wauquier, M. (2020). What's in an agent?. *Morphology*, 30(3), 185-218.

Lipka, Leonhard. 1992. Lexicalization and institutionalization in English and German. *Linguistica Pragensia/Akademie Ved CR, Ústav pro Jazyk Ceskỳ* 1–13.

Mikolov, Tomas, Kai Chan, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In

Références

Proceedings of international conference onlearning representations (iclr). Scottsdale.

Roché, Michel. 2004. Mot construit ? Mot non construit ? Quelques réflexions à partir des dérivés en -ier(e). *Verbum* 26(2). 459–480.

Urieli, Assaf, 2013. *Robust French syntax analysis : reconciling statistical methods and linguistic knowledge in the Talismane toolkit*. Thèse de doctorat. Université Toulouse Le Mirail.

Varvara, Rossella, Gabriella Lapesa & Sebastian Padó. 2021. Grounding semantic transparency in context. *Morphology* 213–234.

Wauquier, M. (2020). *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. Thèse de doctorat en sciences du langage, Université Toulouse Jean Jaurès.

Wauquier, M., Hathout, N. & Fabre, C. (2020). Semantic discrimination of technicality in French nominalizations. *Zeitschrift für Wortbildung/Journal of Word Formation*, 4(2), 100-119.